Modelling Cardiac Patient Set Residuals Using Rough Sets

Aleksander Øhrn* Staal Vinterbo* Piotr Szymański‡ Jan Komorowski*

*Knowledge Systems Group
Dept. of Computer and Information Science
Norwegian University of Science and Technology
N-7034 Trondheim, Norway
e-mail: {aleks, staalv, janko}@idi.ntnu.no

‡Postgraduate Medical School
Dept. of Cardiology
51/59 Grenadierów Str.
01-073 Warsaw, Poland

Abstract

Many medical studies deal with the assessment of the prognostic or diagnostic power of some particular test with respect to some particular medical condition. However, even though a test is deemed to be powerful in this respect, the test may not be strictly needed to perform for everyone. If the test is costly or invasive, this issue is of particular interest. This paper presents a methodology based on rough set theory and Boolean reasoning that can be used to identify those patients for whom performing the test is redundant or superfluous. Furthermore, the methodology enables one to automatically construct a set of descriptive and minimal if-then rules that model the patient group in need of the test. A reanalysis of a previously published real-world dataset of patients with chest pain [1] is used as a case study.

INTRODUCTION

Exercise testing provides important diagnostic and prognostic information in patients with known or suspected coronary artery disease. However, a large portion of patients with chest pain may not be able to exercise adequately, thus reducing the detection of coronary artery disease. For such patients, alternative stress modalities have to be used. In a recently published study [1], a group of patients with chest pain underwent a dobutamine-atropine technetium-99m sestamibi single-photon emission computed tomography (SPECT) scintigraphic study. Using multivariate logistic regression, it was found that the single most important independent predictor for future hard cardiac events (cardiac death or nonfatal myocardial infarction) was an abnormal scan pattern. However, performing a scintigraphic scan is a relatively expensive procedure, and may for some patients not really be fully necessary as knowledge of the outcome of the scan may be redundant with respect to making a prognosis. If one through considering combinations of more readily available parameters could make the same decisions, one could thereby minimize the number of scans acquired and hence potentially cut both costs and use of resources. This paper proposes a procedure for doing this, using elements from rough set theory [3] coupled with Boolean reasoning techniques [2].

Rough set theory deals with the approximation of sets, e.g. the set of all patients that will either die or have a myocardial infarction within a certain follow-up period, or the set of all patients susceptible to a certain treatment. The same group of patients as in [1] have been reanalyzed using a rough set approach. In an identification step, the framework is used to pinpoint exactly which patients where knowledge of the outcome of the scan is strictly required with regards to predicting future hard cardiac events. Furthermore, in a subsequent modelling step one can extract minimal if-then decision rules that describe how to identify this patient group.

PRELIMINARIES

An *information system* is a pair $\mathbb{A} = (U, A)$, where U is a non-empty finite set called the *universe* and A is a non-empty finite set of *attributes* such that $a : U \to V_a$ for every $a \in A$. The set V_a is called the *value set* of a. A *decision table* is any information system of the form $\mathbb{A} = (U, A \cup \{d\})$, where $d \notin A$ is a distinguished attribute called the *decision attribute*. The elements of A are called *condition attributes*.

Let $\mathbb{A} = (U, A)$ be an information system. For any $B \subseteq A$ is associated an equivalence relation $IND(B)$ as defined below. $IND(B)$ is called the *B-indiscernibility relation*.

$$IND(B) = \{(x, x') \in U^2 \mid \forall a \in B,\ a(x) = a(x')\}$$

1091—8280/97/$5.00

A *reduct* is a minimal set of attributes $B \subseteq A$ such that $IND(B) = IND(A)$, i.e. a minimal attribute subset that preserves the partitioning of the universe. The set of all reducts of an information system $\mathbb{A}$ is denoted $RED(\mathbb{A})$.

Let $\mathbb{A}$ be information system with n objects. The *discernibility matrix* of $\mathbb{A}$ is a symmetric $n \times n$ matrix with entries c_{ij} as given below.

$$c_{ij} = \{a \in A \mid a(x_i) \neq a(x_j)\} \text{ for } i, j = 1, ..., n$$

A *discernibility function* $f_{\mathbb{A}}$ for an information system $\mathbb{A}$ is a Boolean function of m Boolean variables $a_1^*, ..., a_m^*$ (corresponding to the attributes $a_1, ..., a_m$) defined as below, where $c_{ij}^* = \{a^* \mid a \in c_{ij}\}$. The set of all prime implicants of $f_{\mathbb{A}}$ determines the set of all reducts of $\mathbb{A}$.

$$f_{\mathbb{A}}(a_1^*, ..., a_m^*) = \bigwedge \left\{ \bigvee c_{ij}^* \mid 1 \leq j \leq i \leq n,\, c_{ij} \neq \emptyset \right\}$$

Let $\mathbb{A} = (U, A)$ be an information system and let $B \subseteq A$ and $X \subseteq U$. We can approximate X using only the information contained in B by constructing the *B-lower* and *B-upper approximations of* X, denoted $\underline{B}X$ and $\overline{B}X$ respectively, as defined under.

$$\underline{B}X = \{x \mid [x]_B \subseteq X\} \quad \overline{B}X = \{x \mid [x]_B \cap X \neq \emptyset\}$$

The objects in $\underline{B}X$ are certain members of X, while the objects in $\overline{B}X$ are possible members of X. The set $\overline{B}X - \underline{B}X$ is called the *B-boundary region of* X, and consists of those objects that we cannot decisively classify into X. The set $U - \overline{B}X$ is called the *B-outside region of* X, and consists of those objects that certainly do not belong to X. A set is said to be *rough* (resp. *crisp*) if the boundary region is non-empty (resp. empty).

The *rough membership function* $\mu_X^B(x)$ defined below [4] quantifies the degree of relative overlap between the set X and the equivalence class to which x belongs. The rough membership function can be interpreted as a frequency-based estimate of $\Pr(x \in X \mid x, B)$.

$$\mu_X^B(x) = \frac{|[x]_B \cap X|}{|[x]_B|} \in [0, 1]$$

The formulas for the lower and upper set approximations can readily be generalized to some arbitrary level of precision $\pi \in [\frac{1}{2}, 1]$ by means of the rough membership function [7], as shown below. Possible ties in the case of $\pi = 0.5$ can be resolved by assigning the objects in question to the interior of the set. Note that the lower and upper approximations as originally formulated are obtained as a special case with $\pi = 1.0$.

$$\underline{B}_\pi X = \{x \mid \mu_X^B(x) \geq \pi\} \quad \overline{B}_\pi X = \{x \mid \mu_X^B(x) > 1 - \pi\}$$

DATA MATERIAL

Table 1 summarizes the contents of the decision table $\mathbb{A} = (U, A \cup \{d\})$. There are 418 objects in the universe U, with the heavily skewed DEATHMI attribute as the decision attribute d and all other attributes defining the set of condition attributes A. All attributes are binary-valued, signifying the absence or presence of some feature. The cut-off values used for discretization of the inherently numerically valued attributes were decided upon externally by medical experts. There were no missing values in the data.

The 418 objects are all patients with chest pain, referred for the evaluation of suspected myocardial ischemia. The data in the decision table is largely the same as having been previously analyzed in [1], with some exceptions. 26 of the 418 patients with early elective coronary revascularization within 60 days after stress testing were excluded from the multivariate statistical analysis performed in [1]. None of these sustained a major cardiac event before coronary revascularization. The 26 patients could not be excluded from the present analysis, due to lack of knowledge of exactly which patients they were.

The attributes listed in Table 1 and used in the present analysis are a subset of those used in [1]. The endpoint is the same, reporting any subsequent hard cardiac events within a certain follow-up period. The APSTRESS and STT attributes were acquired by means of the dobutamine-atropine stress test, while the SCANABN data originates from the scintigraphic scan. All other attributes were selected primarily due to their simplicity and ease of acquisition. For a more detailed description of the patient group and the precise semantics of each attribute, see [1].

METHODOLOGY

Identification

Let $X = \{x \mid \text{DEATHMI}(x) = 1\}$, and let $\mathbb{A} = (U, A)$ refer to the decision table described in the previous section. The set X thus defines the patients who either died or had a heart attack within the follow-up period. In the identification stage, rough approximations of X are calculated using various attribute sets of interest. Of interest is then to identify which patients that, when selected attributes (most notably the SCANABN attribute) are excluded, migrate into the boundary region from either the lower approximation or the outside region. This corresponds to those patients where knowledge of the removed attributes is strictly required for identifiability. For all other patients, acquiring the additional knowledge will have no effect with respect to the classification into X. Either because this knowledge

Attribute	Description	% of 0s	% of 1s	Attribute	Description	% of 0s	% of 1s
AGE	Over 70 years old?	80.9%	19.1%	GENDER	Male?	43.1%	56.9%
OLDMI	Prior infarction?	51.0%	49.0%	HFMED	History of dec. cordis?	80.4%	19.6%
HYPERT	Hypertension?	57.2%	42.8%	ANGP	History of angina?	75.1%	24.9%
DM	Diabetes?	85.6%	14.4%	APSTRESS	Angina during stress?	72.2%	27.8%
SMOK	Smoking?	72.0%	28.0%	STT	ST-T changes?	68.9%	31.1%
CHOL	Hypercholesterolemia?	74.9%	25.1%	SCANABN	Abnormal scan?	31.6%	68.4%
DEATHMI	Cardiac death or inf.?	88.8%	11.2%				

Table 1: Distribution summary of the 418 patients in decision table $\mathbb{A}$.

is superfluous (i.e., we already have enough information to classify them within a reasonable degree of certainty), or because it won't help us anyway (i.e., we still won't be able to classify them within a reasonable degree of certainty even if we perform the additional tests). Figure 1 displays this graphically. The grayed area denotes the set of migrating patients, when attributes $C \subseteq B$ are removed from $B \subseteq A$.

The rough approximations can be made with varying degrees of precision π. To evaluate the approximations, we define *sensitivity*, *specificity* and *accuracy* as the following quantities for any $B \subseteq A$:

$$\begin{aligned}
\text{Sensitivity}(B,\pi,X) &= \frac{|\underline{B}_\pi X \cap X|}{|X|} \\
\text{Specificity}(B,\pi,X) &= \frac{|(U-\overline{B}_\pi X)\cap(U-X)|}{|U-X|} \\
\text{Accuracy}(B,\pi,X) &= \frac{|X|}{|U|}\text{Sensitivity}(B,\pi,X) + \\
&\quad \frac{|U-X|}{|U|}\text{Specificity}(B,\pi,X)
\end{aligned}$$

Modelling

Let A be as previously defined and let $C \subseteq B \subseteq A$. The set of migrating patients can formally be defined as follows:

$$\begin{aligned}
\text{Boundary}(B,\pi,X) &= \overline{B}_\pi X - \underline{B}_\pi X \\
\text{Migrate}(B,C,\pi,X) &= \text{Boundary}(B-C,\pi,X)\cap \\
&\quad (U - \text{Boundary}(B,\pi,X))
\end{aligned}$$

If a patient is a member of the migration set, this means that we cannot classify the patient within a reasonable degree of certainty without additional knowledge that might help discern him/her. Relating to Figure 1, this constitutes the union of the grayed areas. In our case, we are especially interested in the instance where $C = \{\text{SCANABN}\} \subseteq B = A$.

Let X as previously denote the set of patients with a future hard cardiac event. Define a decision table $\mathbb{A}'_\pi$ with condition attributes $A - \{\text{SCANABN}\}$, and a new decision attribute d'_π defined as:

$$d'_\pi(x) = \begin{cases} \text{Yes} & \text{if } x \in \text{Migrate}(A, \{\text{SCANABN}\}, \pi, X) \\ \text{No} & \text{otherwise} \end{cases}$$

The set $X'_\pi = \{x \mid d'_\pi(x) = \text{Yes}\}$ obviously constitutes the set of patients for whom knowledge of the SCANABN attribute is strictly needed in order to classify them. A model of this set in the form of minimal if-then rules can be automatically extracted from $\mathbb{A}'_\pi$ using several techniques [5]. One straightforward scheme is the following, typically employed in a cross-validation setting:

1. Compute the set of reducts $RED(\mathbb{A}'_\pi)$.
2. Filter away "weak" reducts.
3. Overlay each reduct over $\mathbb{A}'_\pi$ to produce a set of decision rules.
4. Filter away "weak" rules.

In practice, it is typically desirable to extract short, approximate rules in order to reveal more general patterns in the data. To this end, dynamic reducts [6] are often computed in the first step. Several criteria may be envisioned in the intermediate filtering steps, e.g. removing reducts/rules with a weak support basis or with a cost associated with them exceeding some threshold.

RESULTS

Table 2 lists the results of performing rough approximations of X (for various levels of precision) with and without some attributes (most notably SCANABN) removed. The numbers indicate the cardinalities of the approximation regions in question. The drops in sensitivity and specificity can be attributed to exactly those patients that migrate into the boundary region when the attributes are removed. Objects in the boundary region are refrained from being classified. Note that a precision level of 0.5 amounts to always selecting the most probable category, and hence by definition results in the empty boundary region.

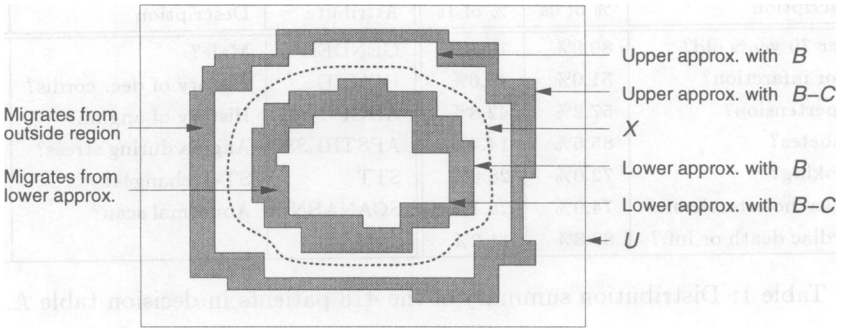

Figure 1: Rough approximations.

Approx. region	$\pi' \in \{1.0, 0.9, 0.8\}$	$\pi' = 0.5$
Upper approx.	35	25
Lower approx.	2	25
Boundary region	33	0
Outside region	383	393
Sensitivity	10.5% (2/19)	84.2% (16/19)
Specificity	96.0% (383/399)	97.7% (390/399)
Accuracy	92.1% (385/418)	97.1% (406/418)

Table 3: Approximating $X'_{1.0}$ with $A - \{\text{SCANABN}\}$ and $\pi' \in \{1.0, 0.9, 0.8, 0.5\}$.

The set X'_π is by definition rough without the SCANABN attribute, and crisp with. One can then proceed to perform a rough metaanalysis of X'_π, for various secondary precision levels π'. Table 3 summarizes a metaanalysis of $X'_{1.0}$.

To get a feel for the type of rules that the rule mining process generates, some actual rules (and their supports/probabilities) from a preliminary extraction session are given below. In real life, the full set of rules would be interpreted by a medical expert.

R_i: OLDMI(0) $\wedge$ HYPERT(0) $\wedge$ GENDER(1)
$\rightarrow$ Migrate(No) [52 objects (1.0)]

R_j: AGE(1) $\wedge$ GENDER(0) $\wedge$ STT(1) $\wedge$ ANGP(0)
$\rightarrow$ Migrate(Yes) [2 objects (1.0)]

R_k: AGE(0) $\wedge$ HYPERT(0) $\wedge$ DM(0) $\wedge$ SMOK(0) $\wedge$ GENDER(1) $\wedge$ APSTRESS(0) $\wedge$ HFMED(0)
$\rightarrow$ Migrate(No) [18 objects (0.818)] $\vee$
Migrate(Yes) [4 objects (0.182)]

DISCUSSION

Since the dataset is not exactly identical as the one used in the statistical analysis in [1], results between the two are not directly comparable. Moreover, the foci of the two analyses are different in nature. Whereas the statistical analysis sought to assess the prognostic power of the nuclear scan procedure, the rough set analysis seeks to identify those patients for whom the procedure will presumably be of no help, and to construct a minimal description of these.

As can be read from Table 2, excluding the SCANABN attribute only results in a minor drop in sensitivity. Since so few patients migrate into the boundary region, this seems to suggest that there is a substantial potential gain in considering combinations of more easily available parameters in lieu of performing a scan. But even though the number of migrating patients is low, what matters in practice is our ability to identify them. If the upper approximation of the set of migrating patients had been extremely large, nothing much would have been gained. However, the upper approximation of the migrating set (even for the most conservative values of π and π') counts no more than 35 people, as can be read from Table 3. In practice, one would presumably send all patients in the upper approximation to acquire a scan. If so, the fact that the lower approximation in Table 3 is so small (and the sensitivity hence so low) may not be an important issue. Also evident from Table 2 is that the drop in both sensitivity and specificity is substantial when *all* stress test information is removed. This confirms the already known fact that stress testing yields valuable prognostic information.

The rough set methodology offers a kind of data compression both in the number of objects (via equivalence classes) and the number of attributes (via reducts), where the degree of loss can be controlled. Furthermore, the automatically extracted if-then rules offer a means of explanation, in contrast to many black-box methods.

Acknowledgements

Thanks to Ron van Domburg at the Thoraxcentre, University Hospital Rotterdam-Dijkzigt, The Netherlands, for making available the data. This work has been supported in part by the EU 4$^{\text{th}}$ Framework project CARDIASSIST and NFR grants #74467/410 and #107409/320.

π	Approx. region	A	$A - \{\text{SCANABN}\}$	$A - \left\{\begin{array}{c}\text{APSTRESS}\\ \text{STT}\\ \text{SCANABN}\end{array}\right\}$
1.0	Upper approx.	78	96	163
	Lower approx.	29	28	19
	Boundary region	49	68	144
	Outside region	340	322	255
	Sensitivity	61.7% (29/47)	59.6% (28/47)	40.4% (19/47)
	Specificity	91.6% (340/371)	86.8% (322/371)	68.7% (255/371)
	Accuracy	88.3% (369/418)	83.7% (350/418)	65.6% (274/418)
0.9	Upper approx.	78	79	92
	Lower approx.	29	28	19
	Boundary region	49	51	73
	Outside region	340	339	326
	Sensitivity	61.7% (29/47)	59.6% (28/47)	40.4% (19/47)
	Specificity	91.6% (340/371)	91.1% (338/371)	86.5% (321/371)
	Accuracy	88.3% (369/418)	87.6% (366/418)	81.3% (340/418)
0.8	Upper approx.	65	67	83
	Lower approx.	29	28	19
	Boundary region	36	39	64
	Outside region	353	351	335
	Sensitivity	61.7% (29/47)	59.6% (28/47)	40.4% (19/47)
	Specificity	94.6% (351/371)	93.8% (348/371)	88.7% (329/371)
	Accuracy	90.9% (380/418)	90.0% (376/418)	83.3% (348/418)
0.5	Upper approx.	52	47	32
	Lower approx.	52	47	32
	Boundary region	0	0	0
	Outside region	366	371	386
	Sensitivity	87.2% (41/47)	80.9% (38/47)	55.3% (26/47)
	Specificity	97.0% (360/371)	97.6% (362/371)	98.4% (365/371)
	Accuracy	95.9% (401/418)	95.7% (400/418)	93.5% (391/418)

Table 2: Approximating X with different attribute sets and $\pi \in \{1.0, 0.9, 0.8, 0.5\}$.

References

[1] M. Geleijnse, A. Elhendy, R. van Domburg et al. (1996), *Prognostic Value of Dobutamine-Atropine Stress Technetium-99m Sestamibi Perfusion Scintigraphy in Patients with Chest Pain*, J. Am. Coll. Card., Vol. 28, No. 2, Aug., pp. 447–454.

[2] F. M. Brown (1990), *Boolean Reasoning*, Kluwer Academic Publishers, Dordrecht.

[3] Z. Pawlak (1991), *Rough Sets - Theoretical Aspects of Reasoning about Data*, Kluwer Academic Publishers, Dordrecht.

[4] Z. Pawlak, A. Skowron (1994), *Rough Membership Functions*, Advances in the Dempster-Shafer Theory of Evidence, R. Yager, M. Fedrizzi and J. Kacprzyk (eds.), Wiley, pp. 251–271.

[5] A. Skowron (1995), *Synthesis of Adaptive Decision Systems from Experimental Data*, Proc. Fifth Scandinavian Conf. on Artificial Intelligence, Trondheim, Norway, May 29–31, Frontiers in Artifical Intelligence and Applications, A. Aamodt and J. Komorowski (eds.), Vol. 28, IOS Press, pp. 220–238.

[6] J. Bazan, A. Skowron, P. Synak (1994), *Dynamic Reducts as a Tool for Extracting Laws from Decision Tables*, Proc. Symp. on Methodologies for Intelligent Systems, Charlotte, NC, USA, Oct. 16–19, Lecture Notes in Artificial Intelligence, Vol. 869, Springer Verlag, pp. 346–355.

[7] W. Ziarko (1993), *Variable Precision Rough Set Model*, J. Comp. Syst. Sci., 46, pp. 39–59.